Journal of Experimental Botany, Vol. 57, No. 7, pp. 1485–1491, 2006
Plant Proteomics Special Issue
doi:10.1093/jxb/erj130 Advance Access publication 21 March, 2006

Journal of Experimental Botany

Arabidopsis thaliana proteomics: from proteome to genome

Sacha Baginsky* and Wilhelm Gruissem

Institute of Plant Sciences, Swiss Federal Institute of Technology, ETH Zürich, Zürich, Switzerland

Received 10 April 2005; Accepted 23 January 2006

Abstract

Proteomics has become an important approach for investigating cellular processes and network functions. Significant improvements have been made during the last few years in technologies for high-throughput proteomics, both at the level of data analysis software and mass spectrometry hardware. As proteomics technologies advance and become more widely accessible, efforts of cataloguing and quantifying full proteomes are underway to complement other genomics approaches, such as RNA and metabolite profiling. Of particular interest is the application of proteome data to improve genome annotation and to include information on post-translational protein modifications with the annotation of the corresponding gene. This type of analysis requires a paradigm shift because amino acid sequences must be assigned to peptides without relying on existing protein databases. In this review, advances and current limitations of full proteome analysis are briefly highlighted using the model plant *Arabidopsis thaliana* as an example. Strategies to identify peptides are also discussed on the basis of MS/MS data in a protein database-independent approach.

Key words: Analysis, *Arabidopsis thaliana*, proteomics.

Introduction

Arabidopsis thaliana has become the plant model organism of choice for which a systems level understanding of complex cellular processes seems to be within reach (Provart and McCourt, 2004). Global analysis of the system components (DNA, RNA, proteins, metabolites) is now possible, although at different analytical depth at present. High-quality genome sequence information is available for *Arabidopsis* (*Arabidopsis* Genome Initiative, 2000), and based on this information GeneChips® have been developed to analyse the transcriptional activity of most predicted genes. The analysis of all proteins (proteome) and all metabolites (metabolome), however, continues to pose significant challenges. Proteins and metabolites are more diverse and biochemically heterogeneous, which precludes the application of a single standardized procedure for their analysis (reviewed in Aebersold and Mann, 2003; Bino *et al*., 2004).

The *Arabidopsis* genome was the first eukaryotic genome that was entirely represented on TILING arrays (Mockler and Ecker, 2005). TILING arrays are high density microarrays of oligonucleotides representing the entire genome. In contrast to GeneChips® they also cover regions that are not predicted to possess coding capacity. Two recent studies reported a number of surprising results from full genome TILING arrays using RNA isolated from cultured *Arabidopsis* cells (Yamada *et al*., 2003; Stolc *et al*., 2005). Both reports showed significant antisense transcription and transcription activity from intergenic regions, suggesting that the transcriptional capacity of the *Arabidopsis* genome exceeds by far current estimates that are based on genome annotation (Yamada *et al*., 2003; Stolc *et al*., 2005). These exciting results raise the question to what extent antisense transcription regulates the translation rate of a mRNA *in vivo* and which of the unexpected transcripts from 'non-annotated' intergenic regions are actually translated into proteins and are not just a consequence of transcriptional read-through. It is clear that a detailed and global analysis of all expressed proteins will help to answer these questions and provide urgently needed complementary information about genome structure, activity, and regulation. Proteomics can also provide information about post-translational protein modifications and subcellular localization of the gene product. This type of information is essential to understand better the complex cellular network, for example, the compartmentalization of

* To whom correspondence should be addressed. E-mail: sacha.baginsky@ipw.biol.ethz.ch

metabolic pathways and the dependencies between RNA, proteins and metabolites.

Proteomics: novel concepts and expected results

Proteomics has a central role in the systems biology workflow and complements the analysis of the transcriptome and the metabolome. For example, it is currently difficult to predict cellular protein concentrations from the abundance of mRNAs, although several reports found positive correlations between transcript levels and the abundance of subsets of all proteins (Gygi *et al*., 1999; Ideker *et al*., 2001; Griffin *et al*., 2002; Washburn *et al*., 2003; Kleffmann *et al*., 2004; Tian *et al*., 2004, reviewed in Greenbaum *et al*., 2003; Hack, 2004). A positive correlation suggests that transcripts are directly translated into proteins. For many pathways, and especially those that are involved in signalling, protein levels can be regulated independently of transcript levels. Such mechanisms include control of the mRNA translation rate or stability of the protein. In addition, antisense transcription probably contributes significantly to the final concentration of a protein (Yamada *et al*., 2003; Stolc *et al*., 2005).

Development of improved high-throughput proteomics techniques has shifted attention to 'protein profiling', which attempts to identify all proteins that are present in a cell. Protein profiling in high-throughput mode is relatively simple and provides a snapshot of the major protein constituents of the cell (reviewed in Aebersold and Mann, 2003; Yates, 2004). Standard protein profiling technologies, however, have two major shortcomings. First, only the most abundant proteins can be detected in a complex protein mixture and, second, the high-throughput analysis is mostly restricted to proteins and peptides in protein databases. Low abundance proteins or proteins translated from alternatively spliced mRNAs that were not predicted correctly, as well as peptides that carry a post-translational modification, will typically escape routine analysis. Despite these drawbacks, protein profiling is an important step towards the systems level analysis of a cell. Desiere and colleagues (2005) suggested using high-throughput proteomics data for the improvement of genome annotation. They used identified, and therefore validated, peptides to improve the genome annotation and to identify splice sites and expressed genome regions that were not predicted to contain protein coding regions. In their analysis, the authors report the unambiguous detection of several SNPs and the confirmation of splice junctions (Desiere *et al*., 2005).

Using proteomics data for genome annotations is preferred over expressed sequence tags (EST) and full coding DNA (cDNA) sequences because annotation can be based on identified peptides. Additional information such as, for example, post-translational modifications can be included directly in the genome annotation. Furthermore, proteome analyses of isolated cell organelles will provide information about the subcellular localization of the gene product. Recent large-scale proteome analyses with *Arabidopsis* chloroplasts (Friso *et al*., 2004; Kleffmann *et al*., 2004), mitochondria (Heazlewood *et al*., 2004), and vacuoles (Carter *et al*., 2004) suggested that the *in silico* prediction of protein localization may not be correct for certain proteins and, therefore, protein localization must be verified by additional experiments. Desiere and colleagues (2005) offer an open source platform, called Peptide Atlas, to allow researchers to submit their MS/MS data and make them available for genome annotation by database curators. The collection of data from different sources is an efficient strategy to increase the coverage of the 'peptide-validated' genome. At the current rate of data collection proteome coverage increases linearly with the number of submitted MS/MS data (Desiere *et al*., 2005), suggesting that there is still a significant gain of information from each experiment. Although it can be expected that the curve of newly discovered peptides will become saturated as more data are being deposited, this stage has not been reached and data collection is still an efficient strategy to increase proteome coverage.

Protein database-independent amino acid sequence determination from MS/MS data

New genome information obtained from peptides that were identified by MS/MS requires a paradigm shift in proteomics. Until now, high-throughput analyses almost exclusively identified those peptides and proteins that are present in protein databases. Thus, it is currently a prerequisite that a gene has been identified and annotated in the genome prior to proteome analysis. This is a dilemma, because new genome information can be expected only from peptides that are not in the database because they might span regions of unusual splice sites, or because they are derived from unpredicted gene regions or they have an unpredicted post-translational modification. Thus, when genome information can be improved from validated peptides, it is necessary and important that the peptide amino acid sequence can be extracted directly from the MS/MS spectrum. In principle, two different approaches can be envisioned that can help to reach this ambitious goal. First, MS/MS data are searched against a protein database and those spectra that were left unassigned despite originating from a true peptide must be identified and subjected to a modified database search. This could be a genome database or a protein database search that includes a variety of different post-translational modifications. This strategy requires a database-independent spectrum scoring scheme to identify putative peptide-derived spectra and to distinguish them from low quality noise fragmentation or contaminants. Second, the amino acid sequence can be extracted directly from the MS/MS spectrum in a fully database-independent fashion. This is

also referred to as '*de novo*' sequencing. Both strategies will be discussed briefly and the available tools for such an analysis will be illustrated.

Searching MS/MS data against genome databases is a time-consuming process that requires enormous computational resources and is, therefore, impracticable for most laboratories. Thus, different strategies were devised to narrow the number of MS/MS spectra that must be analysed in such a way. MS/MS spectra are typically scored by heuristic quality parameters to identify those MS/MS spectra that were derived from true peptide fragmentation (Moore *et al*., 2000; Tabb *et al*., 2003; Bern *et al*., 2004; Purvine *et al*., 2004; Hansen *et al*., 2005; Nesvizhskii *et al*., 2005; Xu *et al*., 2005). When this is combined with a normal database search, those spectra that were left unassigned by the database search despite a high quality are ideal candidates for modified database searches (Nesvizhskii *et al*., 2005). This way only a very small subset of the MS/MS spectra must be subjected to a modified database search, therefore allowing for higher search spaces (the time constraint is kept minimal) and for more careful data analysis, including manual data interpretation.

The first step towards database-independent spectrum scoring usually includes the extraction from true peptide MS/MS spectra of those parameters that are characteristic and indicative of peptide fragmentation in collision-induced dissociation (CID). Several publications have reported such parameters, which are typically dealing with peak height distribution (i.e. intensities of signal peaks compared to noise), overall signal to noise ratio, number of complementing peaks that give rise to the measured parent mass when summed up (i.e. potential b- and y-ions), isotope distribution, number of neutral losses in MS/MS spectra, and the occurrence of amino acid tags. Several implementations of spectra scoring have been released during the last years, assessing all or a subset of the above depicted parameters that are characteristic for true peptide fragmentation (Moore *et al*., 2000; Tabb *et al*., 2003; Bern *et al*., 2004; Purvine *et al*., 2004; Nesvizhskii *et al*., 2005; Xu *et al*., 2005). A discriminant function is calculated to distinguish between 'good' and 'poor' quality spectra, resulting in an estimate of spectral quality. Only those spectra that are of high quality are then used in a modified database search.

In addition to modified database searches, peptide sequences from an MS/MS spectrum can be determined *de novo* in a fully database-independent fashion. *De novo* sequencing tools exclusively use the information in the MS/MS spectrum to derive an amino acid sequence. Several tools are available and some of them achieve a good quality *de novo* sequencing result when applied to high quality spectra (Chen *et al*., 2001; Johnson and Taylor, 2002; Ma *et al*., 2003; Searle *et al*., 2004; Zhang, 2004; Fischer *et al*., 2005; Frank and Pevzner, 2005; Grossmann *et al*., 2005). All tools, however, suffer from inherent difficulties with MS/MS spectra that result from inaccurate measurements, missing peaks (gaps), and chemical or instrument noise. Currently employed software tools can provide a probability whether the extracted amino acid sequence is correct and, in addition, assign probabilities to sequence sub-strings. The best performing tools use probabilistic approaches, for example, PepNovo (Frank and Pevzner, 2005), PEAKS (Ma *et al*., 2003), and an HMM-based implementation (Fischer *et al*., 2005). PepNovo is publicly available and uses a probabilistic network whose structure reflects the physico-chemical characteristics of peptide fragmentation in CID. PEAKS is a commercial software that computes the best possible peptide sequence for a MS/MS spectrum and provides confidence scores for amino acids in the sequence. The HMM defines a generative model to provide emission probabilities for the suggested amino acid sequence of the observed spectrum. Together, the last few years have seen an enormous improvement in mass spectrometry methods and, particularly, in the software tools to aid the MS/MS spectrum analysis. These parallel developments have now paved the way for the use of high-throughput proteomics data to derive highly reliable, high quality peptide assignments that can be used for genome annotation.

Proteome coverage: current limitations and remedies

Important and novel insights from proteomics data for genome annotation require that proteome analysis reaches a satisfactory depth to enable the annotation of large parts of the genome with validated peptides. To date, most large-scale proteome analyses with *Arabidopsis* were performed with isolated organelles, membrane systems, or subcellular structures (reviewed in Peck, 2005). These studies include chloroplasts (Peltier *et al*., 2002, 2004; Ferro *et al*., 2003; Froehlich *et al*., 2003; Friso *et al*., 2004; Huber *et al*., 2004; Kleffmann *et al*., 2004; Baginsky *et al*., 2005; reviewed in Baginsky and Gruissem, 2004; van Wijk, 2004), mitochondria (Brugiere *et al*., 2004; Lister *et al*., 2004; Heazlewood *et al*., 2004,; Millar *et al*., 2005), peroxisomes (Fukao *et al*., 2003), vacuoles (Carter *et al*., 2004; Shimaoka *et al*., 2004), the plasma membrane (Alexandersson *et al*., 2004; Ephritikine *et al*., 2004; Sazuka *et al*., 2004; Marmagne *et al*., 2004; Borner *et al*., 2005) the cell wall (Chivasa *et al*., 2002; Borderies *et al*., 2003; Boudart *et al*., 2005), and cytosolic ribosomes (Chang *et al*., 2005). In general, all studies using subcellular organelles reported the identification of proteins that were not predicted to localize to the organelle under investigation when *in silico* localization prediction tools were used as a benchmark. This suggests that intracellular protein trafficking is more complex than anticipated and that unexpected import routes might exist. It is clear, however, that at the current state of proteome analyses additional experiments are necessary to validate the subcellular localization of proteins. Proteomics data

provide an excellent starting point for the design of such experiments.

Most large-scale proteomics analyses in *Arabidopsis* were designed such that they generally do not allow the identification of post-translational modifications or unusual peptides. Exceptions are those studies that specifically analysed particular post-translational protein modifications. In these instances, the chemical characteristics of the post-translationally modified peptides or proteins were used for their enrichment and subsequent analysis by MS/MS. This strategy has been successfully employed for the large-scale detection of phosphoproteins from the plasma membrane of *Arabidopsis* cultured cells. Phosphopeptides were enriched by metal ion affinity chromatography (IMAC) and analysed by MS/MS analysis, which provided information about the phosphorylation site. This comprehensive type of information was used for the assembly of a phosphorylation site database (Nuehse *et al*., 2003, 2004). Another example is the selective isolation of GPI-anchored proteins from the plasma membrane. Here, membrane vesicles were enriched and GPI-anchored proteins specifically released from the membrane by phosphatidylinositol phospholipase C treatment (Borner *et al*., 2003; Elortza *et al*., 2003). Further examples for the directed analysis of specific post-translational modifications are *S*-nitrosylated proteins (Lindermayr *et al*., 2005) and redox proteomics based on the search for thioredoxin targets (Marchand *et al*., 2004).

The above examples illustrate that a targeted analysis of post-translational modifications is possible. With this type of analysis, however, unexpected post-translational modifications cannot be detected. The lack of comprehensive information is also a shortcoming of routine identification of proteins and peptides. It compounds the problem of reaching high proteome coverage from currently reported proteome information. A large-scale study reported by Giavalisco *et al*. (2005) was designed to achieve complete proteome coverage of *Arabidopsis* cells using 2D gel electrophoresis and MALDI-TOF peptide mass fingerprinting. Although the authors used different tissues to broaden the range of protein fractions in order to increase the probability of detecting different proteins, they found only 663 different proteins that originated from 2943 spots. This number of proteins is surprisingly low considering the number of expected proteins in a cell. All subcellular proteome analyses reported to date identified primarily the most highly abundant proteins, sometimes despite sophisticated protein fractionation strategies (Baginsky *et al*., 2005). Therefore it must be questioned how deep proteome analyses are in practice and what measures can be taken to increase proteome coverage.

Important reasons why certain proteins or peptides escape detection during LC-MS/MS analyses are not just an unusual peptide structure or unanticipated post-translational modifications, but also the effect of 'undersampling' during LC-MS/MS experiments when performed with complex protein mixtures. Liu and colleagues (2004) developed a statistical model, which predicts that as many as 10 LC-MS/MS runs are necessary to reach 95% coverage of all peptides that are theoretically detectable in the mixture. Although runs were performed under the same conditions, each additional run resulted in new peptide identifications (Liu *et al*., 2004). Therefore it seems necessary to run a single peptide fraction several times under the same conditions. The 'undersampling' effect can now be reduced by increasing the scan speed with the new generation of ion trap MS/MS instruments. These 'linear traps' contain a mass analyser with an optimized geometry that allows speeding up the duty cycle of each scan considerably. This way, the scan rate is increased, which results in a significant increase in instrument sensitivity and, therefore, allows low abundance peptides to be detected as well. Although the mass spectrometer hardware continues to be improved to achieve higher sensitivities of mass measurements and thus higher proteome coverage, certain limits will remain.

One of the basic limitations is the dynamic range of protein concentrations in a cell, which exceeds the sensitivity of every mass spectrometry device available to date. The currently most-advanced mass spectrometers can handle a dynamic range of 3–5 orders of magnitude. The dynamic range of protein concentrations in a cell, however, is several orders of magnitude higher. Thus, fractionating proteins or peptides prior to the mass spectrometric analysis is necessary to reduce the dynamic range limitations and to increase proteome coverage. New fractionation tools are continuously being developed and it is not within the scope of this review to highlight all of them. In general, multidimensional protein or peptide fractionation (MudPIT) is a preferred strategy to reduce sample complexity prior to analysis of complex mixtures by MS/MS and has recently been expanded to include additional dimensions to increase proteome coverage (Washburn *et al*., 2001). In general, reducing the complexity and dynamic range of protein samples prior to analysis is perhaps the most promising and efficient strategy to achieve full (or nearly full) proteome coverage. This strategy should include different fractionation techniques such as MudPIT, 2D PAGE, or free flow electrophoresis together with protein fractions from different tissues or subcellular compartments.

Analysis of a complete proteome remains a challenge despite significant advances in mass spectrometry technology and peptide fractionation tools. Such a challenge can best be tackled by a community effort. Integration of data from different sources will increase the information to expand proteome coverage. Genome annotation based on peptide identification in particular requires an open source platform to collect and integrate MS/MS data. The Peptide Atlas platform (Desiere *et al*., 2005) has the potential to develop into such an open source platform that will also serve the *Arabidopsis* community. Many laboratories have already uploaded their data in the SBEAMS database

(Systems Biology Experiment Analysis Management System) which is offered by the Institute of Systems Biology in Seattle (www.sbeams.org) and represents a general framework for systems biology experiments. But the open source platform strategy can provide reliable results only if the shared data have the same high quality and reliability. This is presently guaranteed by a standardized data analysis pipeline, which includes statistical tools such as PeptideProphet (Keller *et al*., 2002; Nesvizhskii *et al*., 2003) to estimate false positive identification rates prior to genome annotation. The authors are convinced that this strategy is most promising for defining proteomes and achieving a genome coverage that is sufficiently deep. The collective proteome information will serve the community to build better tools for functional genomics and to increase our understanding of biological systems.

Acknowledgements

We apologize to all colleagues whose papers were not cited in this review due to space constraints. The authors would like to thank Johannes Fütterer for critical reading of the manuscript and Franz Roos and Jonas Grossmann for help with the literature research. The authors' work is funded by the ETH and the Velux foundation.

References

Aebersold R, Mann M. 2003. Mass spectrometry-based proteomics. *Nature* **422,** 198–207.

Alexandersson E, Saalbach G, Larsson C, Kjellbom P. 2004. *Arabidopsis* plasma membrane proteomics identifies components of transport, signal transduction and membrane trafficking. *Plant Cell Physiology* **45,** 1543–1556.

***Arabidopsis* Genome Initiative.** 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815.

Baginsky S, Gruissem W. 2004. Chloroplast proteomics: potentials and challenges. *Journal of Experimental Botany* **55,** 1213–1220.

Baginsky S, Kleffmann T, von Zychlinski A, Gruissem W. 2005. Analysis of shotgun proteomics and RNA profiling data from *Arabidopsis thaliana* chloroplasts. *Journal of Proteome Research* **4,** 637–640.

Bern M, Goldberg D, McDonald WH, Yates III JR. 2004. Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics* **20,** Supplement 1, 49–54.

Bino RJ, Hall RD, Fiehn O, *et al.* 2004. Potential of metabolomics as a functional genomics tool. *Trends in Plant Sciences* **9,** 418–425.

Borderies G, Jamet E, Lafitte C, Rossignol M, Jauneau A, Boudart G, Monsarrat B, Esquerre-Tugaye MT, Boudet A, Pont-Lezica R. 2003. Proteomics of loosely bound cell wall proteins of *Arabidopsis thaliana* cell suspension cultures: a critical analysis. *Electrophoresis* **24,** 3421–3432.

Borner GH, Sherrier DJ, Weimar T, Michaelson LV, Hawkins ND, Macaskill A, Napier JA, Beale MH, Lilley KS, Dupree P. 2005. Analysis of detergent-resistant membranes in *Arabidopsis*: evidence for plasma membrane lipid rafts. *Plant Physiology* **137,** 104–116.

Borner GHH, Lilley KS, Stevens TJ, Dupree P. 2003. Identification of glycosylphosphatidylinositol-anchored proteins in *Arabidopsis*. A proteomic and genomic analysis. *Plant Physiology* **132,** 568–577.

Boudart G, Jamet E, Rossignol M, Lafitte C, Borderies G, Jauneau A, Esquerre-Tugaye MT, Pont-Lezica R. 2005. Cell wall proteins in apoplastic fluids of *Arabidopsis thaliana* rosettes: identification by mass spectrometry and bioinformatics. *Proteomics* **5,** 212–221.

Brugiere S, Kowalski S, Ferro M, *et al.* 2004. The hydrophobic proteome of mitochondrial membranes from *Arabidopsis* cell suspensions. *Phytochemistry* **65,** 1693–1707.

Carter C, Pan S, Zouhar J, Avila EL, Girke T, Raikhel NV. 2004. The vegetative vacuole proteome of *Arabidopsis thaliana* reveals predicted and unexpected proteins. *The Plant Cell* **16,** 3285–3303.

Chang IF, Szick-Miranda K, Pan S, Bailey-Serres J. 2005. Proteomic characterization of evolutionarily conserved and variable proteins of *Arabidopsis* cytosolic ribosomes. *Plant Physiology* **137,** 848–862.

Chen T, Kao MY, Tepel M, Rush J, Church GM. 2001. A dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology* **8,** 325–337.

Chivasa S, Ndimba BK, Simon WJ, Robertson D, Yu XL, Knox JP, Bolwell P, Slabas AR. 2002. Proteomic analysis of the *Arabidopsis thaliana* cell wall. *Electrophoresis* **23,** 1754–65.

Desiere F, Deutsch EW, Nesvizhskii AI, *et al.* 2005. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biology* **6,** R9.

Elortza F, Nuhse TS, Foster LJ, Stensballe A, Peck SC, Jensen ON. 2003. Proteomic analysis of glycosylphosphatidylinositol-anchored membrane proteins. *Molecular and Cellular Proteomics* **2,** 1261–1270.

Ephritikhine G, Ferro M, Rolland N. 2004. Plant membrane proteomics. *Plant Physiology and Biochemistry* **42,** 943–962.

Ferro M, Salvi D, Brugiere S, Miras S, Kowalski S, Louwagie M, Garin J, Joyard J, Rolland N. 2003. Proteomics of the chloroplast envelope membranes from *Arabidopsis thaliana*. *Molecular and Cellular Proteomics* **2,** 325–345.

Fischer B, Roth V, Roos FF, Grossmann J, Baginsky S, Widmayer P, Gruissem W, Buhmann JM. 2005. NovoHMM: a hidden Markov model for *de novo* peptide sequencing. *Analytical Chemistry* **77,** 7265–7273.

Frank A, Pevzner P. 2005. PepNovo: *de novo* peptide sequencing via probabilistic network modeling. *Analytical Chemistry* **77,** 964–973.

Friso G, Giacomelli L, Ytterberg AJ, Peltier JB, Rudella A, Sun Q, Wijk KJ. 2004. In-depth analysis of the thylakoid membrane proteome of *Arabidopsis thaliana* chloroplasts: new proteins, new functions, and a plastid proteome database. *The Plant Cell* **16,** 478–499.

Froehlich JE, Wilkerson CG, Ray WK, McAndrew RS, Osteryoung KW, Gage DA, Phinney BS. 2003. Proteomic study of the *Arabidopsis thaliana* chloroplastic envelope membrane utilizing alternatives to traditional two-dimensional electrophoresis. *Journal of Proteome Research* **2,** 413–425.

Fukao Y, Hayashi M, Nishimura M. 2002. Proteomic analysis of leaf peroxisomal proteins in greening cotyledons of *Arabidopsis thaliana*. *Plant Cell Physiology* **43,** 689–696.

Giavalisco P, Nordhoff E, Kreitler T, Kloppel KD, Lehrach H, Klose J, Gobom J. 2005. Proteome analysis of *Arabidopsis thaliana* by two-dimensional gel electrophoresis and matrix-assisted laser desorption/ionisation-time of flight mass spectrometry. *Proteomics* **5,** 1902–1913.

Greenbaum D, Colangelo C, Williams K, Gerstein M. 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology* **4,** 117.

Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, Aebersold R. 2002. Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Molecular and Cellular Proteomics* **1,** 323–333.

Grossmann J, Roos FF, Cieliebak M, Jacob R, Lipták Z, Mathis LK, Müller M, Widmayer P, Gruissem W, Baginsky S. 2005. AUDENS: a tool for automated peptide *de novo* sequencing. *Journal of Proteome Research* **4,** 1768–1774.

Gygi SP, Rochon Y, Franza BR, Aebersold R. 1999. Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology* **19,** 1720–1730.

Hack CJ. 2004. Integrated transcriptome and proteome data: the challenges ahead. *Briefings in Functional Genomic and Proteomic* **3,** 212–219.

Hansen BT, Davey SW, Ham AJ, Liebler DC. 2005. P-Mod: an algorithm and software to map modifications to peptide sequences using tandem MS data. *Journal of Proteome Research* **4,** 358–368.

Heazlewood JL, Tonti-Filippini JS, Gout AM, Day DA, Whelan J, Millar AH. 2004. Experimental analysis of the *Arabidopsis* mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins. *The Plant Cell* **16,** 241–256.

Huber CG, Walcher W, Timperio AM, Troiani S, Porceddu A, Zolla L. 2004. Multidimensional proteomic analysis of photosynthetic membrane proteins by liquid extraction-ultracentrifugation-liquid chromatography-mass spectrometry. *Proteomics* **4,** 3909–3920.

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292,** 929–934.

Johnson RS, Taylor JA. 2002. Searching sequence databases via *de novo* peptide sequencing by tandem mass spectrometry. *Molecular Biotechnology* **22,** 301–315.

Keller A, Nesvizhskii AI, Kolker E, Aebersold R. 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry* **74,** 5383–5392.

Kleffmann T, Russenberger D, von Zychlinski A, Christopher W, Sjolander K, Gruissem W, Baginsky S. 2004. The *Arabidopsis thaliana* chloroplast proteome reveals pathway abundance and novel protein functions. *Current Biology* **14,** 362–375.

Lindermayr C, Saalbach G, Durner J. 2005. Proteomic identification of *S*-nitrosylated proteins in *Arabidopsis*. *Plant Physiology* **137,** 921–930.

Lister R, Chew O, Lee MN, Heazlewood JL, Clifton R, Parker KL, Millar AH, Whelan J. 2004. A transcriptomic and proteomic characterization of the *Arabidopsis* mitochondrial protein import apparatus and its response to mitochondrial dysfunction. *Plant Physiology* **134,** 777–789.

Liu H, Sadygov RG, Yates III JR. 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical Chemistry* **76,** 4193–4201.

Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G. 2003. PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry* **17,** 2337–2342.

Marchand C, Le Marechal P, Meyer Y, Miginiac-Maslow M, Issakidis-Bourguet E, Decottignies P. 2004. New targets of *Arabidopsis* thioredoxins revealed by proteomic analysis. *Proteomics* **4,** 2696–2706.

Marmagne A, Rouet MA, Ferro M, Rolland N, Alcon C, Joyard J, Garin J, Barbier-Brygoo H, Ephritikhine G. 2004. Identification of new intrinsic proteins in *Arabidopsis* plasma membrane proteome. *Molecular and Cellular Proteomics* **3,** 675–691.

Millar AH, Heazlewood JL, Kristensen BK, Braun HP, Moller IM. 2005. The plant mitochondrial proteome. *Trends in Plant Sciences* **10,** 36–43.

Mockler TC, Ecker JR. 2005. Applications of DNA tiling arrays for whole-genome analysis. *Genomics* **85,** 1–15.

Moore RE, Young MK, Lee TD. 2000. Method for screening peptide fragment ion mass spectra prior to database searching. *Journal of the American Society for Mass Spectrometry* **11,** 422–426.

Nesvizhskii A, Keller A, Kolker E, Aebersold R. 2003. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry* **75,** 4646–4658.

Nesvizhskii A, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R. 2005. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Molecular and Cellular Proteomics* (Epub ahead of print.)

Nuehse TS, Stensballe A, Jensen ON, Peck SC. 2003. Large-scale analysis of *in vivo* phosphorylated membrane proteins by immobilized metal ion affinity chromatography and mass spectrometry. *Molecular and Cellular Proteomics* **2,** 1234–1243.

Nuehse TS, Stensballe A, Jensen ON, Peck SC. 2004. Phosphoproteomics of the *Arabidopsis* plasma membrane and a new phosphorylation site database. *The Plant Cell* **16,** 2394–2405.

Peck SC. 2005. Update on proteomics in *Arabidopsis*: Where do we go from here? *Plant Physiology* **138,** 591–599.

Peltier JB, Emanuelsson O, Kalume DE, *et al.* 2002. Central functions of the lumenal and peripheral thylakoid proteome of *Arabidopsis* determined by experimentation and genome-wide prediction. *The Plant Cell* **14,** 211–236.

Peltier JB, Ytterberg AJ, Sun Q, van Wijk KJ. 2004. New functions of the thylakoid membrane proteome of *Arabidopsis thaliana* revealed by a simple, fast, and versatile fractionation strategy. *Journal of Biological Chemistry* **279,** 49367–49383.

Provart NJ, McCourt P. 2004. Systems approaches to understanding cell signaling and gene regulation. *Current Opinions in Plant Biology* **7,** 605–609.

Purvine S, Kolker N, Kolker E. 2004. Spectral quality assessment for high-throughput tandem mass spectrometry proteomics. *OMICS: a Journal of Integrative Biology* **8,** 255–265.

Sazuka T, Keta S, Shiratake K, Yamaki S, Shibata D. 2004. A proteomic approach to identification of transmembrane proteins and membrane-anchored proteins of *Arabidopsis thaliana* by peptide sequencing. *DNA Research* **11,** 101–113.

Searle BC, Dasari S, Turner M, Reddy AP, Choi D, Wilmarth PA, McCormack AL, David LL, Nagalla SR. 2004. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS *de novo* sequencing results. *Analytical Chemistry* **76,** 2220–2230.

Shimaoka T, Ohnishi M, Sazuka T, Mitsuhashi N, Hara-Nishimura I, Shimazaki K, Maeshima M, Yokota A, Tomizawa K, Mimura T. 2004. Isolation of intact vacuoles and proteomic analysis of tonoplast from suspension-cultured cells of *Arabidopsis thaliana*. *Plant Cell Physiology* **45,** 672–683.

Stolc V, Samanta MP, Tongprasit W, *et al.* 2005. Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. *Proceedings of the National Academy of Sciences, USA* **102,** 4453–4458.

Tabb DL, Saraf A, Yates III JR. 2003. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Analytical Chemistry* **75,** 6415–6421.

Tian Q, Stepaniants SB, Mao M, *et al.* 2004. Integrated genomic and proteomic analyses of gene expression in mammalian cells. *Molecular and Cellular Proteomics* **3,** 960–969.

van Wijk KJ. 2004. Plastid proteomics. *Plant Physiology and Biochemistry* **42,** 963–977.

Washburn MP, Koller A, Oshiro G, Ulaszek RR, Plouffe D, Deciu C, Winzeler E, Yates III JR. 2003. Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences, USA* **100,** 3107–3112.

Washburn MP, Wolters D, Yates III JR. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology* **19,** 242–247.

Xu M, Geer LY, Bryant SH, Roth JS, Kowalak JA, Maynard DM, Markey SP. 2005. Assessing data quality of peptide mass spectra obtained by quadrupole ion trap mass spectrometry. *Journal of Proteome Research* **4,** 300–305.

Yamada K, Lim J, Dale JM, *et al.* 2003. Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302,** 842–846.

Yates III JR. 2004. Mass spectral analysis in proteomics. *Annual Review of Biophysics and Biomolecular Structure* **33,** 297–316.

Zhang Z. 2004. *De novo* peptide sequencing based on a divide-and-conquer algorithm and peptide tandem spectrum simulation. *Analytical Chemistry* **76,** 6374–6383.